ETS GRE

Assess Ability. Predict Performance.

ETS GRE® Board Research Report
ETS GRE®–17-02
ETS RR–17-05

Evidence of the Generalization and Construct Representation Inferences for the *GRE*® revised General Test Sentence Equivalence Item Type

Isaac I. Bejar

Paul D. Deane

Michael Flor

Jing Chen

March 2017

The report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

****

****

GRE-ETS

PO Box 6000

Princeton, NJ 08541-6000

USA

GRE Board Research Report Series and ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Evidence of the Generalization and Construct Representation Inferences for the *GRE*® revised General Test Sentence Equivalence Item Type

Isaac I. Bejar, Paul D. Deane, Michael Flor, & Jing Chen

Educational Testing Service, Princeton, NJ

The report is the first systematic evaluation of the sentence equivalence item type introduced by the *GRE*® revised General Test. We adopt a validity framework to guide our investigation based on Kane's approach to validation whereby a hierarchy of inferences that should be documented to support score meaning and interpretation is evaluated. We present evidence relevant to the generalization inference as well as evidence of construct representation. We analyzed the pool of sentence equivalence items in three studies. The first and second studies focused on the generalization inference and sought to document the construction principles behind the sentence equivalence items, specifically the nature of the vocabulary tested. The third study focused on construct representation and evaluated the contribution of the stem, the keys, and the distractors to item difficulty. We concluded that the vocabulary tested by the sentence equivalence items is appropriate given the purpose of the GRE, namely, to assist in the selection of graduate students. The difficulty of the items was shown to be, in part, a function of the familiarity of the vocabulary as well as the context in which the vocabulary is tested, which we argue is positive validity evidence.

Keywords Construct representation; vocabulary; difficult modeling validity argument; generalization inference; sentence equivalence items; GRE

doi:10.1002/ets2.12134

The *GRE*® revised General Test was introduced in 2011, after several years of research. As Wendler and Bridgeman (2014) have noted, "[C]hanges in the test-taking population, the relationship of question types to the skills being measured, or expanding on the use of the test scores requires that a careful examination of the test be undertaken" (p. 0.1.1). The purpose of this report is to provide a careful examination of a new item type that is part of the GRE Verbal measure, known as *sentence equivalence*. Performance on the sentence equivalence items, together with performance on *reading comprehension* items, the major component of the verbal measure, and the *text completion* items makes up performance on the verbal measure. The sentence equivalence item type appears to have been developed specifically for the GRE revised General Test, motivated by the limitations of previous item types, especially under adaptive testing-on-demand conditions (Briel & Michel, 2014). The results presented herein represent the first evaluation of the item type following the introduction of the GRE revised General Test.

We adopt a validity framework to guide our investigation. Kane (2006) described a hierarchy of inferences that should be documented to support score meaning and interpretation. At a minimum, these include the scoring, generalization, extrapolation, and decision inferences. In an earlier study (Robin, Bejar, Liang, & Rijmen, 2016), evidence in support of the scoring inference was provided by documenting the fit of the unidimensional item response model used to scale the difficulty and discrimination of the quantitative and verbal GRE measures introduced in 2011. This report focuses solely on the verbal measure and has two aims, namely, providing evidence supporting the generalization inference and evidence supporting construct representation, albeit limited to the portion of the verbal measure based on sentence equivalence items.[1]

One aspect of the generalization inference is concerned with the sampling of content. For example, are the items that appear on a particular version of the test equally representative of some well-defined universe of items? A full evaluation of the generalization inference requires multiple sources of evidence; it is beyond the scope of this report to provide such a comprehensive evaluation of the inference. The earlier study (Robin et al., 2016), although focusing on the scoring

Corresponding author: I. I. Bejar, E-mail: ibejar@ets.org

Table 1 Overview of Studies

		Inference	
Question	Study	Generalization	Construct representation
Can the items in the pool be described by a design pattern that defines the universe of items?	1	√	
Are the words used in the sentence equivalence item type words expected to have been acquired by prospective graduate students regardless of major?	2	√	√
Can the variability in item difficulty be accounted for by the item attributes that reflect a theoretical sound process?	3		√

inference, provided some positive evidence regarding generalization by analyzing data from two points in time based on data obtained during the launch of the revised GRE and data from a year later showing that unidimensionality held. In this report, we focus on a different aspect of the generalization inference, namely, the nature of the universe of items. We consider two aspects, the structure and the content of the items, as described in more detail later. We also evaluate evidence of construct representation of the verbal scores. Within Kane's approach, evidence construct representation is viewed as relevant when the interpretation of scores relies on theoretical constructs, such as process models of the response process. As described subsequently, we rely on a response process model to account for the variability in psychometric attributes of sentence equivalence items.

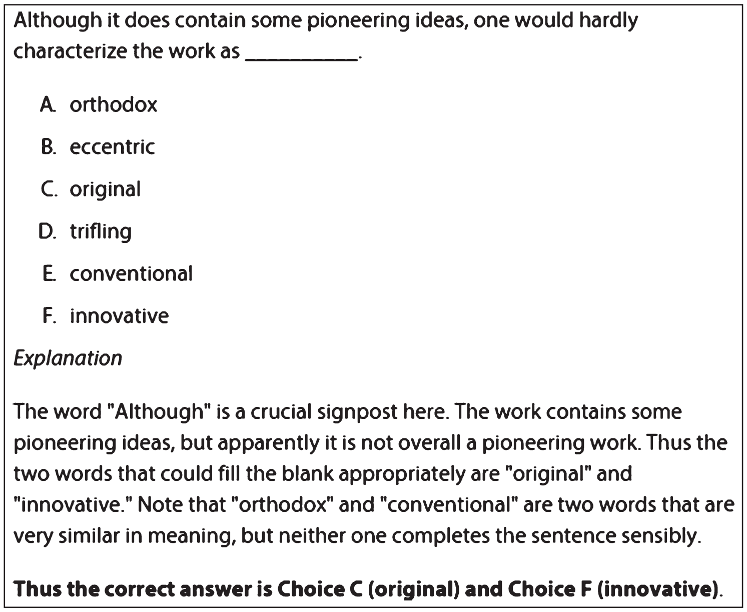
In short, we hope to present evidence that contributes to a validity argument of GRE Verbal scores by documenting the makeup of the universe of sentence equivalence items and by providing a better understanding of the variability in the psychometric attributes of the items.

Research Questions and Overview

Table 1 shows the studies we present in this report and their relationship to the generalization and construct representation inferences. Study 1 aims to document the principles for constructing sentence equivalence items. The goal of Study 2 is to document the nature of the vocabulary used in sentence equivalence items. The nature of the vocabulary is relevant to the generalization inference, because the vocabulary together with the other construction principles defines the universe of items.

We also view that documentation as relevant to construct representation, as indicated in Table 1, because vocabulary is acquired, during reading, for example, by means of psychological processes (McKeown & Curtis, 1987). Study 3 is devoted to the third question we examine, namely, construct representation. The idea of construct representation was proposed by Embretson (1983) as a means of supplementing construct validation, which has historically emphasized covariation evidence, that is, how scores are related, or not related, to each other, including predictive relationships between scores and relevant criteria. Construct representation looks inward, so to speak, to understand the possible mental processes that could underlie scores and whether they are the intended ones. Evidence to that effect, together with evidence that construct-irrelevant processes are not significantly present, would enhance a validity argument and would complement other evidence, such as predictive validity, by suggesting that such predictive relationships are positive for defensible reasons.

The report is organized as follows: We briefly describe the sentence equivalence item type next and the subset of the items we analyzed as part of the three studies that follow. We then present Study 1, which addresses the design of sentence equivalence items. Study 2, which addresses the nature of the vocabulary used by sentence equivalence items, is presented next. Together, the design of the items and the vocabulary used for producing them provide relevant data for arguing the extent to which the portions of the verbal measure consisting of sentence equivalence items support the generalization inference. The results of Study 3 are presented next to address evidence of construct representation through the evaluation of a difficulty model. The last sections address limitations of the studies, and the conclusions we propose based on the three studies.



Although it does contain some pioneering ideas, one would hardly characterize the work as ___________.

A. orthodox

B. eccentric

C. original

D. trifling

E. conventional

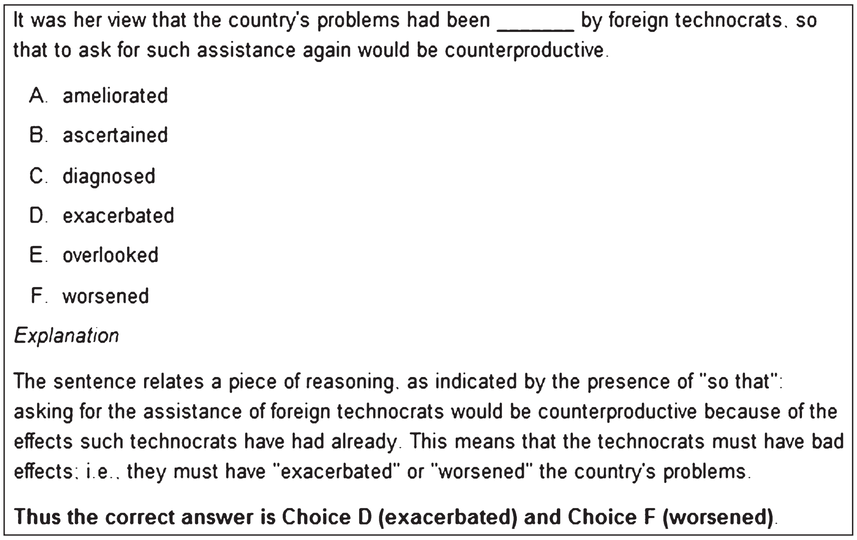
F. innovative

Explanation

The word "Although" is a crucial signpost here. The work contains some pioneering ideas, but apparently it is not overall a pioneering work. Thus the two words that could fill the blank appropriately are "original" and "innovative." Note that "orthodox" and "conventional" are two words that are very similar in meaning, but neither one completes the sentence sensibly.

Thus the correct answer is Choice C (original) and Choice F (innovative).

Figure 1 Sample sentence equivalence item type (adjective).



It was her view that the country's problems had been ________ by foreign technocrats, so that to ask for such assistance again would be counterproductive.

A. ameliorated

B. ascertained

C. diagnosed

D. exacerbated

E. overlooked

F. worsened

Explanation

The sentence relates a piece of reasoning, as indicated by the presence of "so that": asking for the assistance of foreign technocrats would be counterproductive because of the effects such technocrats have had already. This means that the technocrats must have bad effects; i.e., they must have "exacerbated" or "worsened" the country's problems.

Thus the correct answer is Choice D (exacerbated) and Choice F (worsened).

Figure 2 Sample sentence equivalence item type (verb).

The Item Pool

We had access to the entire sentence completion item type for this study, including the item response theory (IRT) operational discrimination and difficulty parameters, *a* and *b* (see Robin et al., 2016). Figures 1 and 2 show sample items.[2] The instructions for answering the items ask the test taker to choose two options:

> Instructions: Select the two answer choices that, when used to complete the sentence, fit the meaning of the sentence as a whole and produce completed sentences that are alike in meaning.

Of the 1,147 items, 346 contain multiword expressions (MWEs), as shown in Table 2. Such items present a challenge for the analysis we perform of the item pool by means of natural language processing (NLP) methods.[3] Excluding MWE items left 800 items for the study. The set of 800, without MWEs, was divided into a development set of 300 randomly chosen items, and the remaining 500 were labeled the cross-validation or test set.

Table 2 Descriptive Statistics of Discrimination and Difficulty for Sentence Equivalence Items Based on Multiword Expression and Single Words

		Discrimination (*a*)		Difficulty (*b*)	
Items	*N*	Mean	*SD*	Mean	*SD*
Multiword expression items	346	0.78	0.25	0.05	0.91
Non-multiword expression items	800	0.81	0.27	0.03	0.89

Study 1

An approach to evaluating the generalization inference is to identify the principles for generating items. Compared to simply evaluating the items in the pool to see if they represent a well-defined universe, an identification of the generating principles has the advantage that is not limited to the items presently in the pool, which makes the evaluation more enduring and applicable to the evolving universe of items. The purpose of the study is to describe the syntactic and semantic attributes of sentence equivalence items.

In Kane's framework, the generalization inference takes different forms depending on the purpose of the assessment, although the ability to create forms containing a representative sample from the universe of potential items applies to any test. There at least two senses of representativeness applicable to sentence equivalence items used in the GRE. One sense of representativeness is with respect to the universe of items: Is there a well-defined universe of items, or the means for creating such a universe, such that when sampling from that universe, it can be assumed that representative and comparable samples will be drawn regardless of when the items were produced? In the case of verbal tests, the other sense of representativeness is with respect to the nature of the vocabulary tested. That is, is the vocabulary used in creating items well defined and appropriate given the purpose of the test? The latter is the subject of Study 2.

An alternative to enumerating all the items to define a universe is to define the construction principles used in writing the items. Within evidence-centered design (Mislevy & Haertel, 2006), design patterns and task models have been proposed as a means to facilitate the conception and creation of items. A design pattern is ideally arrived at early in the development of an assessment and identifies knowledge, skills, and abilities (KSAs) as well as the relevant contexts for eliciting such KSAs. Design patterns evolve into a set of task models from which items are actually produced. Mislevy (2011) saw design patterns as the basis for streamlining the production of test content:

> Supporting tools such as the design patterns … [enable] developers [to] think through the assessment argument without getting tangled up in the details of implementation. Generative schemas for families of tasks are important for assessments that need to *generate multiple forms*. (p. 6, emphasis added)

In practice, design patterns are instantiated by a set of more concrete specifications (item or task models) from which items are then produced. Such task models for GRE Verbal items have been retrospectively inferred for earlier versions of the GRE (Deane & Sheehan, 2003; Sheehan, Kostin, & Futagi, 2005; Sheehan & Mislevy, 2001). Similarly, Bejar, Chaffin, and Embretson (1991) retrospectively identified the design pattern and corresponding item models underlying analogy items in a previous version of the GRE.

The rationale for the elimination of antonyms and analogy items in the previous GRE version led to consideration of several item types that assess vocabulary effectively and efficiently. The sentence equivalence emerged as the winner. Among its virtues is the fact that the probability of guessing the correct answer is lower compared to a standard multiple-choice item, by virtue of the fact that the correct answer requires two choices. Of course, creating such items is not simply a matter of including two keys among the choices. For an item that calls for two correct answers to be effective, it needs to be designed carefully. For example, if the two keys were standard synonyms, it would be possible to answer the items without referring to the stem by simply identifying which pair of options are synonyms, which would have defeated the purpose of moving away from the antonym and analogy items, which were thought of as decontextualized vocabulary item types. Instead, the two keys in a sentence equivalence item are chosen to be synonymous in the context of the stem. That is, the design intention is to make the stem an important component of the item answering process. Similarly, the distractors are also important, of course, in the sense that they do not give away the keys by, for example, being totally unrelated among

Table 3 Frequency, Discrimination, and Difficulty of the 800 Items

		Discrimination (*a*)		Difficulty (*b*)	
Part of speech	*N*	Mean	*SD*	Mean	*SD*
Adjectives	364	0.78	0.28	0.10	0.88
Adverbs	13	0.91	0.31	−0.38	1.01
Nouns	206	0.85	0.27	0.10	0.81
Verbs	216	0.81	0.25	−0.13	0.84
Total	799	0.81	0.27	0.03	0.86

Note. Discrimination and difficulty correspond to the *a* and *b* parameter estimates on a two-parameter logistic item response theory model. The part of speech was not available for one item.

themselves or to the key. Thus the goal of the item type is to engage the test takers in a lexical task not unlike an authentic reading process where the exact meaning of the text must be arrived at from multiple cues, as in normal reading.

We examined two construction principles based on the development sample, namely, the part of speech (POS) the item is based on and the semantic class(es) called for by the stem. Neither of these is part of the operational tags or metadata kept for each item; they were obtained as part of this study. The POS of the sentence equivalence item is determined by the syntactic role of the blank. In the examples given, Figure 1 contains an item classified as "adjective," whereas Figure 2 contains an item classified as "verb." The keys and distractors would then follow the same POS.

Table 3 shows the frequencies, means, and standard deviations of difficulty and discrimination for the items in the 800 items. As can be seen, most items are based on adjectives, followed by verbs and nouns. There are relatively few items based on adverbs. As can be seen, with the exception of items based on adverbs, which are somewhat easier and more discriminating, items based on adjectives, nouns, and verbs are equally difficult and discriminating.

A second item attribute we consider is the semantic class(es) called for by the stem. The syntactic and semantic nature of sentence-based items has been found to be related to difficulty. For example, Bejar, Stabler, and Camp (1987) found that syntactic complexity was related to difficulty of sentence *correction* items for the Test of Written English (TSE). Similarly, Sheehan and Mislevy (2001) found that the semantic nature of the sentence *completion* GRE item type used in the previous version of the GRE was a source of difficulty.

During the redesign of the GRE and the conception of the sentence equivalence item type, several task models corresponding to different semantic templates were considered, although not necessarily as a basis for manipulating difficulty. The five task models considered were *synonymous*, *definitional*, *antonymical*, *local context*, and *global context*. Figure 3 is taken from an unpublished internal development document[4] and illustrates each of these categories.

The rationale for choosing these specific task models was not explicitly given, although it is consistent with the semantic nature of an earlier sentence-based item type (Sheehan & Mislevy, 2001). Here we focus on describing the difficulty and discrimination of items based on these task models. (We present additional relevant information as part of Study 2 when we analyze vocabulary used in the stems.) However, because the task model is not coded operationally, and obtaining it for all the items would be costly, we conducted a pilot study based on 100 randomly chosen items to determine the extent to which the different task models were present and to evaluate the cost of classifying all the items in the pool.

For that purpose, a lead test developer for the sentence equivalence item type classified a random set of items. The test developer first studied the taxonomy in Table 3 and coded the 100 items. The results are shown in Table 4. Three of the items had been deactivated and could not be found for purposes of this study. The bulk of the items fall into three categories: definitional – synonymic, definitional – antonymic, and context – local. With respect to discrimination, the notable finding is that the items classified as context – local and context – global have lower discrimination on average. With respect to difficulty, the definitional – antonymic item appears to be the easiest, whereas the items classified as both synonymic and antonymic appear to be hardest, although there were relatively few items in that category.

As noted earlier, classifying the items according to the taxonomy is not part of the operational process of creating and maintaining the items. To establish the utility of the taxonomy, a much larger sample of items would need to be studied, and it would need to be established that independent coders agree substantially on the classification of items.[5]

Until such a study becomes possible, and based on the small sample of items, it appears that an approach to creating more difficult items, which tend to be in higher demand, is to write items that are classified as synonymic – antonymic.

Definitional: The sentence supplies explicit information about what is missing. Hypotheses about what is missing can be firmly established.	
Synonymous (a synonym or definition of the missing term is given or implied)	In the 1950s the nation's inhabitants were __________: **most of them knew very little about foreign countries**. [insular / provincial] The macromolecule RNA is **common to all living beings**, and DNA, which is found in all organisms except some bacteria, is almost as __________. [universal / ubiquitous] Ever a **demanding** reader of the fiction of others, the novelist Chase was likewise often the object of __________ analyses by his contemporaries. [exacting / meticulous]
Antonymical (an antonym of the missing term or a definition of the missing term's opposite is given or implied)	The Chancellor **castigated** one regime for its mistreatment of political opponents, **whereas** he inexplicably __________ another that had committed similar acts. [exculpated / absolved] Even though the governor was __________ the arguments in favor of the proposal for the new highway construction, he **nevertheless decided to veto** the proposal. [well-disposed toward / sympathetic to] **While** in many ways their personalities could not have been more different—she was ebullient where he was glum, relaxed where he was awkward, **garrulous** where he was ________—they were surprisingly well suited. [laconic / taciturn]
Both (contains both antonymical and synonymous information about the missing term)	The artist's apparently **casual, improvisatory** pictures are not the _______ renditions they may seem: her pictures are, in fact, the result of a **highly disciplined style**. [offhand / cavalier]
Contextual: The sentence supplies only implicit clues about what is missing. Hypotheses about what is missing are likely to be less specific, if they are formed at all.	
Local context (words or phrases in the sentence point in the direction of the missing term, e.g., whether it is positive or negative)	**Overlarge, uneven, and ultimately disappointing**, the retrospective exhibition seems **too much like special pleading for a forgotten painter of real but** __________ talents. [limited / circumscribed] According to some political analysts, the candidate's **occasionally rambling responses to questions suggest** that he has been out of circulation for a while and **his debating skills need to be** _______. [honed / enhanced] The plan, which the engineers said would **save the aquifer by reducing pumping** to _______ levels, has **passed a governmental environmental review** but faces opposition from outdoor and environmental groups. [innocuous / benign]
Global context (the sentence as a whole may yield some information about what is missing, but no specific words or phrases are especially helpful)	Newspapers report that the former executive has been trying to keep a low profile since his ________ exit from the company. [indecorous / unseemly]

Figure 3 Description of sentence equivalence task models.

Table 4 Frequency, Discrimination, and Difficulty of the 100 Items in the Development Set by Task Model

		Discrimination (*a*)		Difficulty (*b*)	
Task model	*N*	Mean	*SD*	Mean	*SD*
Deactivated	3	0.91	0.45	−0.12	0.95
Definitional – synonymic	28	0.85	0.23	0.11	0.88
Definitional – antonymic	24	0.90	0.32	−0.28	0.62
Both synonymic and antonymic	5	0.80	0.14	0.38	0.72
Context – local	35	0.74	0.25	0.10	0.90
Context – global	5	0.75	0.11	0.03	0.80
Total	100	0.82	0.26	0.02	0.82

Finally, we describe the length of the stem. Intuitively, the longer a sentence is, the more it may require working memory capacity (Just & Carpenter, 1992) or introduce more complex semantic relations (Andrews, Birney, & Halford, 2006). However, because the stem in the case of sentence equivalence items provides the context for the blank, sentence length can also have a facilitative effect by providing more hints about the blank. In the end, it is an empirical question what the role of length is on item difficulty, because the item writer may use length as an item writing design variable. For example, if the word to be tested is infrequent, the item writer might consciously test in the context of a longer stem that would contain relevant clues regarding the blank, as opposed to testing in the context of a short, complex sentence.[6]

Figure 4 shows the distribution of stem length for the 800 items. As can be seen, stems range from 10 to 50 words, although the bulk of the items are from 20 to 40 words. Figure 5 shows the scatterplot of length and difficulty; the correlation is −.08. That is, items with longer stems tend to be easier, although the relationship is very tenuous; it is significant at a .05 significance level.

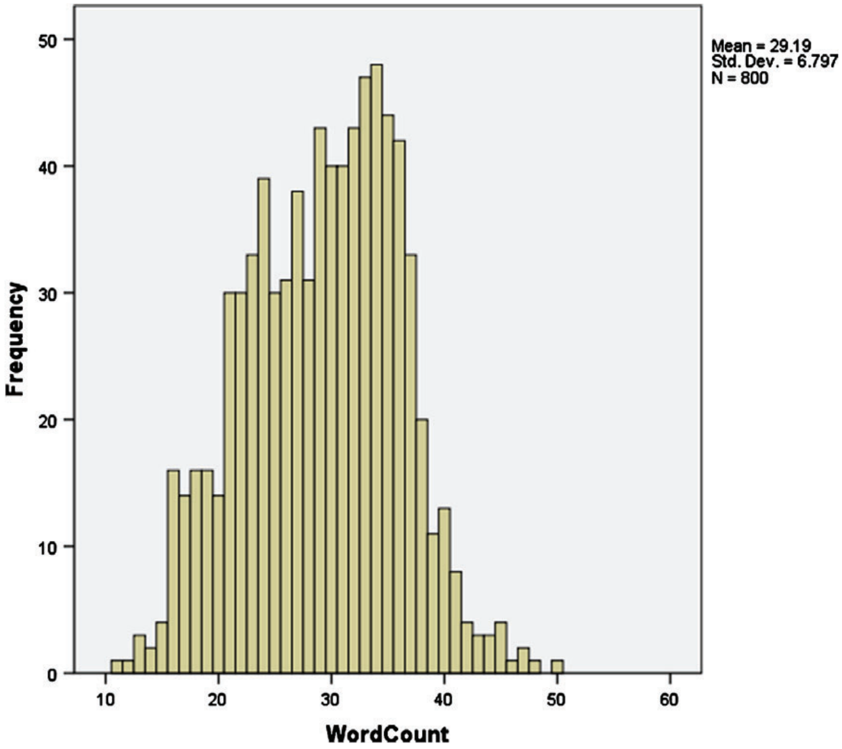


Figure 4 Distribution of stem length for 800 items in the development set.

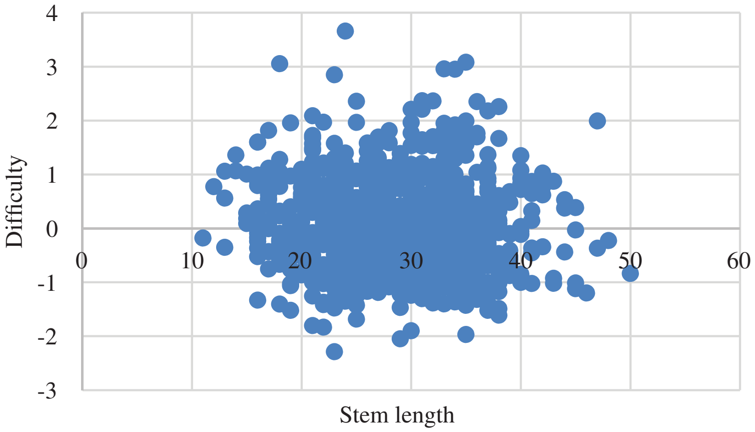


Figure 5 Scatterplot of stem length and item response theory difficulty parameter *b*.

Summary of Study 1

Study 1's goal was to better understand the universe of sentence equivalence items. We described two item characteristics, the syntactic role of the blank in the stem and the semantic categorization of the items, as well as the length of the stem. We noted that in nearly one half of the items, the blank is an adjective, whereas adverbs are rarely used. We also discovered that taxonomy of task models had been formulated during the redesign of the GRE, describing the nature of the semantic relations in the stem. Because the task models are not used explicitly in the development of new items, as part of this study, an experienced test developer classified a random sample of 100 items. Although no firm conclusions can be drawn based on a small sample size, it appears that items classified as antonymic and synonymic are harder. Finally, items classified as context–global or context–local appear to be somewhat less discriminating. If these findings could be corroborated in a future study, it could help to solve the practical problem of producing hard items that are also discriminating. We also examined stem length and its relationship to difficulty and found it to be slightly negative, suggesting that, other things being equal, longer stems are associated with easier items.

Study 2

The goal of this substudy is to provide a detailed analysis of the vocabulary used by sentence equivalence items. Such analysis informs the generalization inference as well as being relevant to construct representation.

According to Burton, Welsh, Kostin, and VanEssen (2009), the rationale for including acquired vocabulary in a verbal reasoning measure is that it

> indicates that critical reading and reasoning have occurred in the past, and it is also an important tool for facilitating future comprehension and expression. A broad vocabulary suggests that a broad array of texts were read in the past and constitutes evidence that a range of subject matter can be learned and understood in the future. (p. 10)

Unlike the traditional vocabulary items previously used by the GRE, the current sentence equivalence item provides a sentence as the context, and it is more consistent with the vocabulary acquisition process. According to Lohman (2000),

> the high correlation between vocabulary knowledge and reasoning seems to reflect primarily the fact that word meanings are generally inferred from the contexts in which they are embedded. But there is a synergism here in that vocabulary knowledge allows comprehension and expression of a broader array of ideas, which in turn facilitate the task of learning new words and concepts. Thus language functions as a vehicle for the expression, refinement, and acquisition of thought. (p. 319)

Hence a vocabulary measure can be justified as proxy for the reading history of a student and the ability to learn future words. Those who have read more have had the opportunity to acquire a larger vocabulary and have honed the skills necessary for acquiring additional vocabulary. Given that graduate education entails reading complex texts, it is reasonable for an admissions examination to include an assessment of the vocabulary a student has acquired.

Nevertheless, it is still necessary to demonstrate that the vocabulary used is appropriate. For a graduate admissions test, it is reasonable for the words to be academic in nature, although it may not be defensible for item difficulty to be dependent on specialized vocabulary that only a subset of the test-taking population would have had an opportunity to learn. As an example, vocabulary found in the topic of Greek mythology could be so specialized that it may be expected that only a fraction of the test takers may have been exposed to it, which could be unfair to the majority of test takers. To address the reasonableness of the vocabulary used in the sentence equivalence items, we obtained the list of all the words used as keys, all those used as distractors, and the words that make up the stem for the 800 items that did not contain MWEs. We investigated the following questions:

Research Question 1. Are the words chosen as keys in sentence equivalence items appropriate for the purpose of assessing readiness for graduate-level work?

Research Question 2. Are the words chosen as distractors in sentence equivalence items generally comparable to the keys, covering a similar vocabulary range?

Research Question 3. Are the words used in sentence equivalence items comparable to the words GRE test takers use in their own writing when they respond to the two prompts that make up the GRE Analytical Writing measure?

Based on the answers to these questions, can we conclude that the vocabulary tested by the GRE is appropriate for the purpose, that is, the vocabulary is such that those applying to graduate school would have had a reasonable opportunity to encounter and acquire? The approach we take to answering these research questions is descriptive. As we show later, the GRE vocabulary is sophisticated enough that the readily lexical resources we had ready access to do not include every GRE word. This increases the complexity of any analysis because the data that are missing are not missing at random.

Methods

One sense of "appropriateness" in the first research question is whether GRE test takers whose native language is English would have had an opportunity to be exposed to and learn the vocabulary used by GRE items.[7] We relied on three measures that can be viewed as indicators of the opportunity to encounter the words. They are described next.

Lexical Measures

Word Frequency

Word frequency is one of the most commonly used metrics for describing vocabulary (Breland, 1996) and has a long standing in psychometric research (Breland, Jones, & Jenkins, 1994; Carroll, 1980; Kirkpatrick & Cureton, 1949). We used standardized frequency indices (SFI)[8] first formulated by Carroll, Davies, and Richman (1971) derived from the Touchstone Applied Science Associates (TASA) corpus (Zeno, Ivens, Koslin, & Zeno, 1995). TASA SFI have been demonstrated to predict text grade level and candidate writing skill in a variety of applications, including Educational Testing Service (ETS; see, e.g., Sheehan, Flor, & Napolitano, 2013). Examination of the words that fall in specific frequency ranges suggests the following interpretation of TASA SFIs: Everyday vocabulary typically falls somewhere above an SFI value of 60 – 80, whereas academic vocabulary typically falls somewhere between an SFI value of 40 and 60, and rarer, more specialized vocabulary typically falls below an SFI value of 35 – 40.

Age of Acquisition

Carroll and White (1973) established the concept of age of acquisition (AofA) as an explanation for frequency and other word difficulty effects, with the hypothesis that earlier-acquired words will be easier to recognize, understand, and recall for expressive use. Carroll and White's studies used a self-reporting methodology to collect estimates of the age at which individuals believed they had learned individual words and demonstrated that these estimates, averaged over multiple individuals, provided a reliable measure of word difficulty. Various efforts have been made to produce larger AofA lexicons and explore their psycholinguistic implications (see, e.g., Morrison, Ellis, & Quinlan, 1992), but the largest effort to date is Kuperman, Stadthagen-Gonzalez, and Brysbaert (2012), who used a crowdsourcing methodology to collect AofA ratings for 30,000 English words. In this study, we use Kuperman et al.'s (2012) AofA ratings. In these ratings, the core vocabulary is judged as having been acquired between ages 4 and 6 years, with almost all words being judged as having been acquired by age 15 years.

Living Word Vocabulary Grade Level

Grade level estimates represent a different approach to characterizing vocabulary. The Living Word Vocabulary (LWV) represents the outcomes of one of the few publicly available studies that systematically measured the grade level at which two thirds of the students in a wide range of US classrooms demonstrated evidence of knowing more than 30,000 word meanings (Dale & O'Rourke, 1976). Students were sampled at Grades 4, 6, 8, 10, and 12, with an additional sample for first-year college (Grade 13) and fourth-year college (Grade 16). The LWV has formed the basis for a variety of follow-up studies, most notably Biemiller and Boote (2006).

To illustrate these three measures, Table 5 shows the TASA, LWV, and AofA measures for the keys and distractors of the two sample items presented earlier. We can see that *ameliorated*, which was used as a distractor, is very infrequent, is classified as a college-level word (Grade 16), and is reported as being acquired at 14 years of age. By comparison, *orthodox* is observed more frequently, is assigned a 10th-grade level, and is reported as having been acquired at 13 years of age.

Table 5 Lexical Measures for the Keys and Distractors of the Sample Items in Figures 1 and 2 (in Alphabetical Order)

Key or distractor	TASA SFI	LWV grade level	Age of acquisition	Used as
Ameliorated	23.8	16	13.67	D
Ascertained	37.7	8	12.68	D
Conventional	49.5	12	9.26	D
Diagnosed	41.9	8	10.72	D
Eccentric	42.7	8	12.6	D
Exacerbated	23.8	13	13.93	K
Innovative[a]	41.1		11.17	K
Original	58.0	6	7.67	K
Orthodox	44.8	10	13.16	D
Overlooked	46.9	6	9.72	D
Trifling	38.9	8	12.05	D
Worsened	39.5	6	8.33	K

Note. LWV = Living World Vocabulary; SFI = standardized frequency indices.
[a]This word did not appear in LWV.

Clearly the scale of these measures needs to be interpreted carefully. Word frequency is an objective measure, but it is highly dependent on the corpus on which it is based, although Breland (1996) showed that the relationships of frequency estimates from different corpora are highly correlated. The LWV is the result of a large study but based on student responses. As with any survey in which the respondent does not have a stake, the issue of motivation and fatigue can potentially determine the outcome (Braun, Kirsch, & Yamamoto, 2011). Finally, with respect to AofA, it is hard to imagine that a respondent actually remembers the age at which he or she acquired a word. It is more likely that given such a task, respondents make an inference as to what is being asked and respond accordingly.

While the scale and uncertainty of these measures suggest caution in interpreting them, they may still be valid indicators of the same construct. For our purposes, the construct of interest is the opportunity to learn the words. The more frequently a word occurs in a corpus that represents a universe of well-defined texts, the more likely it is that a student will encounter that word. In the case of AofA and LWV, the intention is to reference the interpretation to age and grade level, respectively, which would be helpful for interpretation purposes. Although one might reasonably question the absolute *referenced interpretation* of age and grade given the data collection methods (averages of the recalled AofA, or whether the word was answered correctly by two thirds of students in a given grade), they may be still indicators, along with word frequency, of the opportunity test takers may have had to acquire the word.

The correlations among the three measures for all unique words from all the 800 items were analyzed by means of principal components analysis. The first factor accounted for 82% of the total variance, suggesting that they in fact have much in common. This finding suggests that these three lexical measures appear to be indicators of the familiarity of words, despite having been collected by very different methods.

One potential problem in using these measures to describe items is that not all words are defined for each measure. The missing words are typically rarer words, but which words are missing differs significantly across the different datasets. This allows some potential for the trends we observe to be less clear for words not present in all features used to measure word familiarity. To address this concern, we also developed a fourth lexical measure, designed to combine information from the other three and recast the multiple measures on a common scale, namely, grade level. To that effect, we used LWV grade levels as the dependent variable and built regressions models in which we entered various combinations of targeted features (TASA SFI, AofA, plus two additional features—frequency in a corpus of student essays written in ETS's *CRITERION*® online writing evaluation service and the square root of word length in characters). Where all features were defined for a word, we used the predicted value from the regression using all features. Where a feature was missing, we imputed a value based on the regression that used the other (available) features for that word. The resulting variable (predicted Living Word grade level) had a Pearson correlation with an LWV grade level of 0.71.

For each analysis presented in the following pages, we will also present the results using predicted Living Word grade level, not so much as a key element of the analysis but to help with interpretation. Table 6 illustrates roughly what kinds of words appear at each of the levels of the predicted grade level variable, which ranges continuously from 0 to 16.

Table 6 Illustration of Predicted Living Word Vocabulary Level

Predicted Living Word grade level	Sample words
0–1	*Asked, birthday, clothes, eating, getting, hand, mom, people, shoes* …
1–2	*Ate, beds, changed, door, eyes, friend, green, hear, it, jump, kitchen, legs* …
2–3	*Afraid, balloon, coffee, driver, early, farmer, grew, hope, indoors, jelly* …
3–4	*Add, belt, channel, disappear, emergency, fight, glue, hate, include, juicy* …
4–5	*Accept, bandage, checkup, dolphin, edge, feature, giggle, hateful, instant* …
5–6	*Accomplish, beggar, centipede, detergent, exchange, film, goggles, healing* …
6–7	*Absorb, bookshelf, champion, denim, exhaust, focus, glimpse, hayloft* …
7–8	*Abundance, betray, collide, dedicate, employer, fraction, glare, husks* …
8–9	*Abolish, bookmobile, combustion, dormant, excusable, furnace, glacial* …
9–10	*Abductor, bespectacled, clamber, dispute, exemplify, footmen, glorify* …
10–11	*Abdicate, beatitudes, centrifugal, demographics, eyewash, factoid, globule* …
11–12	*Aberrant, bedcover, causality, demystify, edify, filigree, glib, heckler* …
12–13	*Abhor, bastion, capricious, defray, emulsify, filibuster, glamor, heretic* …
13–14	*Abrogate, bastions, carapace, desiccate, efficacious, fiduciary, golem* …
14–15	*Absinthe, bursar, cacophony, doxology, empiricist, fecund, gamut, hypoxia* …
15–16	*Antebellum, bibulous, coadjustor, dross, emesis, geodetic, hirsute, ischemia* …

Features of words used as keys (Research Question 1). We examine the distribution of words used as keys by

- determining the values of all three summary measures for each word used as a key;
- calculating the minimum and maximum values of each summary measure, by individual item;
- calculating descriptive statistics for the overall distribution and for the distribution of minimum and maximum values; and
- examining histograms that show the shapes of these distributions.

Features of words used as distractors (Research Question 2). We apply the same methods to the distractors. We expect the distribution of distractors to be similar to the distribution of keys, although because there are four distractors and only two keys, we anticipate somewhat greater variability in the properties of distractors.

Features of words used in the item stem (Research Question 3). The same methods are applied to the words used in the stem. However, because the stems simulate texts, we expect to observe distributions that contain more common words. To further assess the comparability of the text that makes up the stems, we compare the corpus of stems with a corpus of 600 student essays written as responses to the GRE issue writing prompt.

Results

Features of Words Used as Keys (Research Question 1)

Each item contained two words defined as keys. In the 800-item sample, 84% of keys appeared in one item, 12.1% appeared in two items, and 3.2% appeared in three items, with six words appearing in four items, one in five items, and one in seven items, for a total of 1,329 distinct words. Ninety-five of these words did not have a TASA SFI value, 127 did not have an estimated AofA, and 201 did not have an LWV grade level.

The majority of keys fell between a TASA SFI of 30 and 50, with a mean of 39.32. The majority of keys were words estimated to have been acquired between 8 and 14 years of age, with the mean AofA falling at 11.51. The majority of keys were words falling between Grades 6 and 13, with a mean LWV grade level value of 10.06 (see Table 7).

We then examined the differences between the pairs of keys used in the same item as shown in Table 8 by calculating the minimum and maximum values for each feature by item. The less frequent keys had a TASA SFI of 34.58, compared to the more frequent keys, which had a TASA SFI of 43.87. The earlier-acquired keys had a mean estimated AofA of 10.27, compared to a mean estimated AofA of 12.76 for the later-acquired keys. The lower grade-level keys had a mean grade level value of 8.30, compared to a grade level value of 11.82 for the higher grade-level keys.

Figure 6 illustrates what these differences mean by showing the grade level distribution of the key pairs. The majority of lower grade-level keys fell in the Grades 4–10 range. By contrast, the majority of higher grade-level keys were at Grade 12 and above. Figure 7 shows the frequency distribution of the harder key in each item. It is worth noting that only a small

Table 7 Descriptive Statistics for the Word Tokens in the Key Pairs Contained in the 800 Items

Feature	*N*	Minimum	Maximum	Mean	*SD*
TASA SFI	1,507	11.50	69.00	39.32	10.10
Mean estimated age of acquisition	1,475	3.48	18.71	11.51	2.72
Living Word Vocabulary grade level	1,401	4.00	16.00	10.06	4.05
Valid *N* (listwise)	1,317				

Note. SFI = standardized frequency index.

Table 8 Descriptive Statistics for the Word Tokens in the Key Pairs Contained in the 800 Items

		Easier key		Harder key	
Feature	*N*	Mean	*SD*	Mean	*SD*
TASA SFI	792	43.87	9.15	34.58	8.76
Age of acquisition	772	10.27	2.54	12.76	2.26
Grade level	790	8.30	3.67	11.82	3.64

Note. SFI = standardized frequency indices.

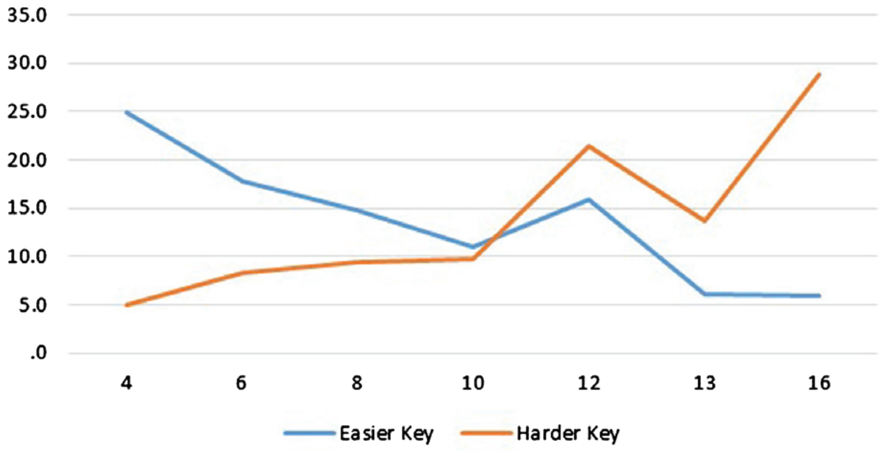


Figure 6 Grade level of easier and harder key.

proportion of the total set of these harder keys is in the lowest portion of the word frequency range, under a TASA SFI of 25, as seen in Figure 7.

There are a relatively small number of words within the TASA dataset with SFIs below 20. Without indicating which of these appear in the GRE set, it is worth noting that they are not, for the most part, incredibly obscure, though they are definitely highly academic or relatively specialized vocabulary, such as *synoptic*, *rectified*, *divulges*, *percolation*, *ineptitude*, *congeals*, *endoscopic*, *foundational*, *sweepstakes*, *nullifying*, *occludes*, *elucidated*, *subdisciplines*, *or overabundance*.

The distribution of harder keys corresponds to the following distribution on the predicted grade level variable (see Figure 8). Because this measure represents the prediction of grade level from other features, it is on a continuous scale (unlike the original LWV grade level estimates, which assigned words to the level at which two thirds of respondents got the question with that item correct). By this measure, the vast majority of the harder words in each pair of keys tend to fall at an estimated grade level between eighth grade and first-year college.

Features of Words Used as Distractors (Research Question 2)

Each of the 800 items contained four distractors; 77.3% of the distractors were used in one item, 15.7% were used in two items, 4.2% were used in three items, and 1.6% were used in four items. A very small number were used more frequently (20 words used five times, 4 words used six times, 2 words used seven times, 1 word used eight times, and 1 word

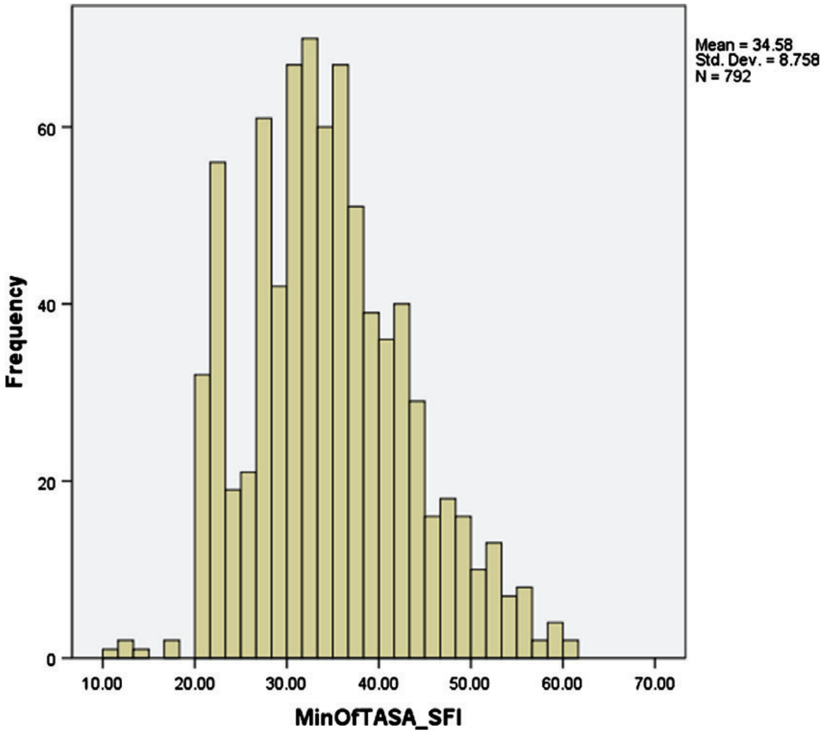


Figure 7 Frequency distribution of harder keys.

used nine times). There were a total of 2,384 distinct distractors, of which 443 had also been used as keys in some other item; 137 of the distractor words were not associated with a TASA SFI value, 167 were not associated with an estimated AofA, and 334 were not associated with a LWV grade level. The descriptive statistics for the distractor sets are shown in Table 9.

Like the keys, the majority of the distractors fell between TASA SFI values of 30 and 50. The mean TASA SFI was 40.59. The majority of distractors had estimated AofA between 8 and 14, with a mean AofA at 11.15. The majority of distractors fell between LWV grade levels 6 and 13, with a mean grade level of 9.41.

We then examined the differences between the easiest and hardest distractors in each key set. The easiest distractors had a mean TASA SFI of 48.72, compared to a mean TASA SFI of 32.03 for the harder distractors. The easiest distractors had a mean estimated AofA of 8.91, compared to a mean estimated AofA of 13.42 for the hardest distractors. The easiest distractors had a mean grade level of 6.41, compared to a mean grade level of 12.67 for the hardest distractors (see Table 10). Figure 9 illustrates these differences by showing the grade level distributions of the easiest and hardest distractors. The easiest distractors range mostly between grade levels 4 and 8, whereas the hardest distractors range mostly at grade level 12 and higher.

Once again, if we examine the distribution of the predicted grade level feature for the harder distractor words in each item, we get a distribution in which the bulk of the harder distractor words range from eighth grade to first-year college (see Figure 10).

Features of Words Used in the Item Stems (Research Question 3)

As can be seen in Table 11, stem words had much higher TASA frequencies than keys or distractors (mean TASA SFI = 65.97), much lower estimated AofA (mean AofA = 6.24), and much lower grade-level estimates (mean LWV grade level = 4.97).

Finally, we compared the vocabulary in the stems of the 800 items with the vocabulary that test takers deploy when responding to the GRE Analytical Writing measure. As can be seen in Table 12, the frequency by predicted grade level was comparable for the stems and the essays. The major difference is that the proportion of fourth- and sixth-grade words is slightly higher in the GRE essays, whereas the stem sentences contain slightly larger percentages of college-level (Grades

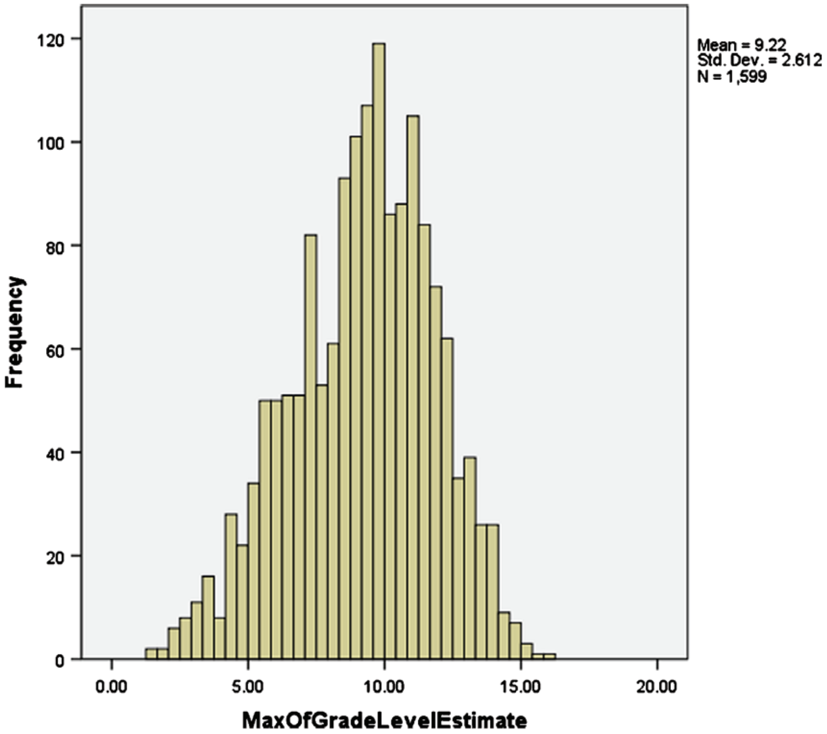


Figure 8 Predicted grade level for the harder key in each item.

Table 9 Descriptive Statistics for the Words in the Distractor Sets Contained in the 800 Items

Feature	*N*	Minimum	Maximum	Mean	*SD*
TASA SFI	3,046	3.50	76.40	40.59	9.43
Mean estimated age of acquisition	3,005	3.89	18.71	11.15	2.52
Living Word Vocabulary grade level	2,791	4.00	16.00	9.41	3.76
Valid *N* (listwise)	2,677				

Note. SFI = standardized frequency indices.

Table 10 Descriptive Statistics for the Easiest and Hardest Words in the Distractor Sets Contained in the 800 Items

	Easiest distractor		Hardest distractor	
Feature	Mean	*SD*	Mean	*SD*
TASA SFI	48.72	7.49	32.03	7.09
Age of acquisition	8.91	2.10	13.42	1.78
Grade level	6.41	2.62	12.67	2.89

Note. SFI = standardized frequency indices.

13 and 16) vocabulary words. However, as can be seen, predicted grade level was not available for 11% of words that appear in the corpus of student essays. Some of these missing words may reflect typos, names, and similar elements rather than a different vocabulary.

Discussion of Study 2

Features of Words Used as Keys (Research Question 1)

The lexical properties of sentence equivalence items appear to be consistent with its design and the purpose of the test.[9] The keys cover the full range of vocabulary from core vocabulary to graduate level. All levels are represented, with an

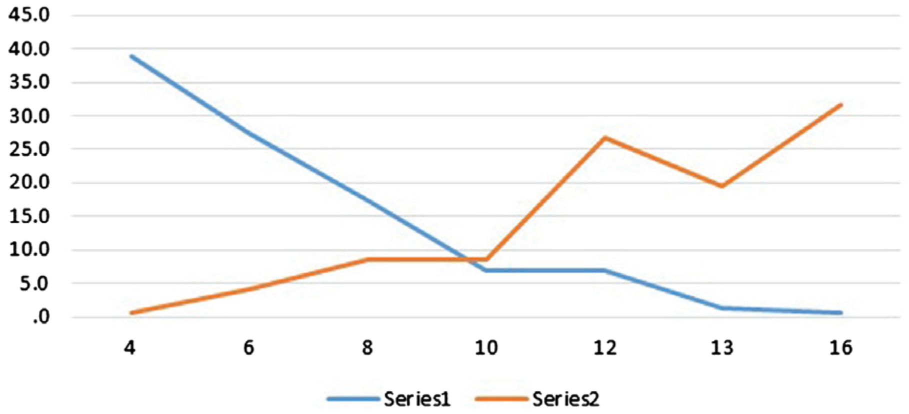


Figure 9 Grade-level distribution of the hardest and easiest distractors.

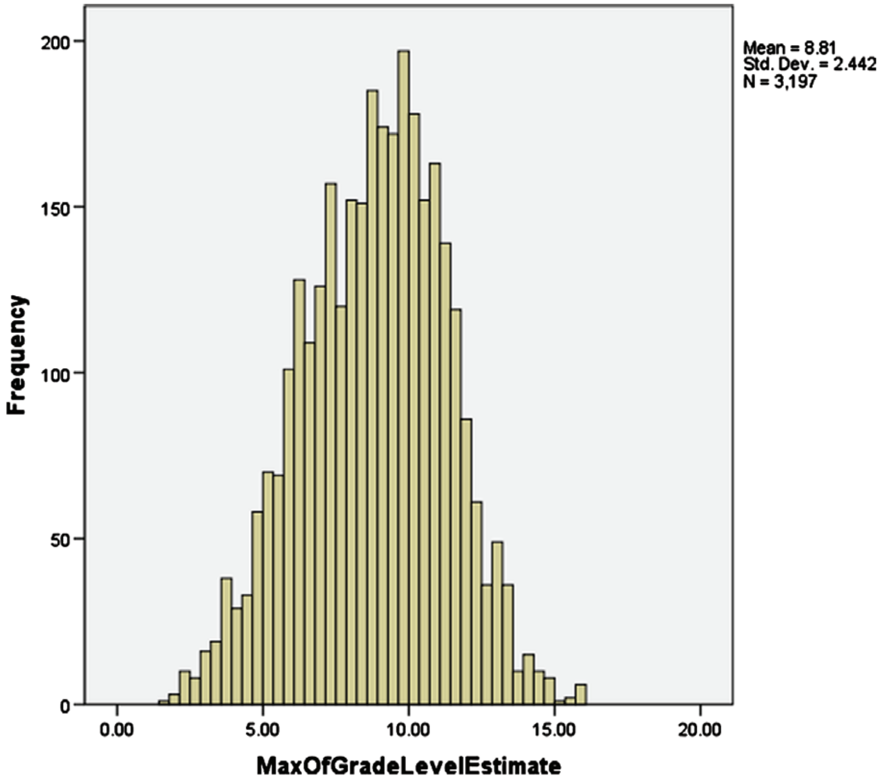


Figure 10 Maximum predicted grade level for all the stems.

emphasis on words that are likely to appear reasonably often in academic texts (e.g., TASA SFIs between 30 and 50, with estimated AofA above age 11 and grade level estimates ranging from high school to college). The distractors display very similar distributions. Overall, the selection of keys and distractors seems to be consistent with a design intended to achieve broad coverage of the vocabulary candidates will need to perform well when reading texts at college and postgraduate levels.

On a more detailed examination, the key pairs appear to be selected so that one of the keys is likely to be significantly easier (more frequent, with a lower AofA and grade level). This effect may be due in large part to the fact that synonyms are unlikely to occur with equal frequency; having selected one synonym that appears in a source sentence, it is quite probable that the second synonym will be less frequent. However, the frequency range of the rarer key is quite reasonable for vocabulary students should have acquired by the time they enter postgraduate studies (TASA mean frequency in the mid-30s, mean estimated AofA near 13 years, and LWV grade levels near early college [Grade 13]).

Table 11 Descriptive Statistics for the Words in the Stems of the 800 Items Analyzed

Feature	*N*	Minimum	Maximum	Mean	*SD*
TASA SFI	22,321	3.50	88.30	65.97	15.41
Mean estimated age of acquisition	21,157	2.37	17.22	6.24	2.63
Living Word Vocabulary grade level	21,636	4.00	16.00	4.97	2.21
Valid *N* (listwise)	20,785				

Note. SFI = standardized frequency index.

Table 12 Percentage of Vocabulary in Each Grade Level, Comparing Stem Sentences to a Random Sample of 600 GRE Issue Essays

Grade level	GRE Essay	Stem sentences
4	66.2	64.9
6	12.7	9.7
8	5.3	4.1
10	2.0	2.2
12	2.0	2.2
13	0.4	0.7
16	0.4	0.8
Total (%)	89	85

Relatively few words appear to fall at the very hard end of the distribution, and those that do appear, from the predicted LWV grade level estimates, to be words that students can reasonably be expected to learn by the end of college. The major threat to this conclusion would be if the words that are missing from the component features (TASA SFI, LWV, or AofA) are particularly hard or rare. The distribution of keys on the predicted grade level feature (which is defined for all words) suggests that vocabulary words not defined for one or more of these features still fall into the expected range (high school to college) when students might reasonably be expected to acquire them for a test like the GRE.

Features of Words Used as Distractors (Research Question 2)

The easiest and hardest distractors are roughly similar to the easier and harder keys, although because they represent extremes of four rather than two words, the easiest and harder distractors are slightly easier, or slightly harder, than the corresponding keys. That is, the mean TASA SFI for the easier key is 43.87, compared to 48.72 for the mean TASA SFI of the easiest distractor, whereas the mean TASA SFI for the harder key is 34.58, compared to a TASA SFI for the hardest distractor of 32.03. Similarly, the mean estimated AofA for the easiest distractor is 8.91, compared to 10.27 for the easier key, whereas the mean estimated AofA for the hardest distractor is 13.42, compared to 12.76 for the harder key. The mean LWV grade level is 12.67 for the hardest distractor, compared to 11.82 for the hardest key. Overall, however, the patterns are similar and suggest (especially considering the very close overall means) that the keys and distractors are drawn from essentially the same pool of vocabulary, which ranges from core oral vocabulary to postgraduate and represents the full range, with an emphasis on rarer, higher grade-level, later-acquired words, though without a large proportion of extremely rare words that would be seldom encountered, even in postgraduate texts. Once again, the distribution on the predicted grade level feature suggests that words missing from any one measure are still words learned in the high school to college span, which is reasonable for words to be tested on a graduate examination.

Features of Words Used in the Item Stems (Research Question 3)

The distribution of words in the stem sentences appears to be very similar (though slightly more challenging) than the distribution of words in student essays produced by GRE candidates and follows a typical word frequency distribution, with a few common words and a larger number of rare words. The observed patterns are consistent with the hypothesis that the stems provide samples of sentence content similar to those that students deploy in their own writing and represent the most compelling evidence that the vocabulary tested by the sentence equivalence item type is appropriate for the test-taking population.

Study 3: Difficulty Modeling

Studies 1 and 2 provided documentation regarding the definition of the universe of sentence equivalence items and whether the vocabulary tested is defensible given the purpose of the test. Study 3 is concerned with an accounting of variability of item difficulty, which, as noted earlier, addresses the response process, specifically, whether the sentence equivalence item type requires a more complex response process than earlier formats used to assess vocabulary. The goal, however, is not to provide a full accounting of difficulty variability but rather to test the presence of components of difficulty: Is the variability in difficulty a function of the nature of the keys, the nature of the distractors, and the nature of the stem? A finding that all three components of the sentence equivalence items contribute independently to difficulty would suggest that students must reason as intended when answering the items, namely, taking into account the context in which the vocabulary question appears, and that the distractors do not give away the correct answer. By contrast, if difficulty were to be a function only of the familiarity about the words representing the key, an alternative conclusion might be that the sentence equivalence item type is not necessarily an improvement upon the item type it replaced because it would imply that it functions as a decontextualized vocabulary item type requiring only word familiarity, as might have been the case with the earlier GRE vocabulary item types. Similarly, if the distractors were not to influence difficulty, it would suggest they play no role in the item solution process. In a worst case scenario, the distractors could give away the intended correct response without requiring the test taker to engage in the deeper process intended. In short, the goal of Study 3 is not necessarily to develop a comprehensive model of difficulty but rather to show that keys, distractors, and the stems are independent contributors to difficulty. Evidence to that effect would be validation that intentions to move away from an approach to measurement vocabulary that is decontextualized are satisfied by the sentence equivalence item type.

Difficulty modeling typically starts with a psychological model relevant to the item type and domain in question (Embretson & Gorin, 2001; Gorin & Embretson, 2006). Given our more modest goal, we postulate a model that focuses on the immediate objective of establishing the independent contribution of the stem, keys, and distractors to item difficulty. Specifically, we view the process of answering sentence equivalence items as an iterative process:

1. A test taker reads the stem and attempts to infer the blank word from the stem.
2. The initial inference is compared to the options. With knowledge that two of the options are keys, pairs of candidate keys are evaluated by the test taker for suitability by evaluating the synonymity of pairs based on the option set and the context provided by the stem. The relationship, or lack thereof, among distractors and between distractors and keys is involved in this process.
3. Once a candidate key pair is tentatively identified, each key is substituted in the blank and the fit of the substitution is evaluated for each candidate key.
4. The process is repeated as necessary.

This simple model engages the three "moving parts" of a sentence equivalence item — stems, keys, and distractors — which we label context, word familiarity, and depth of familiarity.

The outcome of the first step is assumed to depend on the nature of the sentence and the nature of the blank. The syntactic and semantic nature of the stem, and the syntactic role of the blank, could make the process of identifying the correct answer easier or harder. The results from Study 1 suggest that items with antonymic–synonymic semantic content appeared to be harder; items where the blank was an adverb appeared to be easier. Also, items with longer stems appeared to be easier. The facilitating effect of length is plausible. For example, longer stems provide more contexts from which to identify the correct answer. These actual and potential associations with difficulty are not necessarily universal, however, and could be specific to the GRE item pool. That is, item writers probably use the different item attributes as design variables. For example, a difficult word may be accompanied by a longer stem to reduce the purely vocabulary load of such an item.

A potentially important item attribute is the nature of the stem text, especially the text surrounding the blank. That is, the words surrounding the blank may cue the blank by virtue of being a frequent sequence of words. Sequences of words are referred to as *n*-grams, where the "*n*" indicates the length of the sequence. Just like words (1-grams) vary in the frequency with which they appear in some corpus of text, *n*-grams also vary in frequency. Although the role of frequency in item difficulty is well established (see later), the frequency of *n*-grams, short sequences of words, is less so. The computation of *n*-gram frequencies requires very large corpora so that their frequency can be estimated. Such corpora and the means to analyze them are a fairly recent development emanating from NLP sciences where language models, essentially the list of *n*-grams and their frequencies, are used as part of the process of speech recognition (Jurafsky & Martin, 2009) and similar NLP tasks. Not surprisingly, the role *n*-grams in item difficulty has not been studied extensively. However, Deane (2003) proposed word fit cosines as a measure of the relationship between words that tend to co-occur. Deane et al. (2014) described word fit cosine as a measure of "how plausible a word sounds in a phrase, based upon corpus data" (p. 14) and found that word fit cosines were related to item difficulty. The psychological plausibility of measures or indicators of item difficulty based on *n*-grams is additionally supported by research showing that subjects' subjective estimates of the frequency of *n*-grams correspond to the frequency in a large corpus (Shaoul, Westbury, & Harald Baayen, 2013) and the responses of subjects to an open-ended cloze task (Shaoul, Harald Baayen, & Westbury, 2014).

Upon reading the stem, word familiarity begins to exert a stronger influence on the item solution process. If the blank is a word that occurs infrequently in print, a test taker would be less likely to identify a word that fits the blank without also examining the possible choices, other things being equal. The frequency of words that appear in vocabulary items is known to be related to item difficulty. For example, Kirkpatrick and Cureton (1949) found the correlation of frequency and difficulty for the Army General Classification Test to be .47. Similarly, Carroll (1980) found high correlations between *SAT*® test difficulty and word frequency. Bejar et al. (1991) also found that word frequency played a role in the difficulty of a subtype of analogy items. It is reasonable to assume that sentence equivalence writers rely on word frequency or familiarity of the keys to some extent as one way to encourage the deeper thought process that sentence equivalence items intend to elicit.

In addition to the keys and the context provided by the stem, the distractors, the four nonkey options, play an important role in the process of answering a sentence equivalence item. Their familiarity, the relationships, or absence thereof among the distractors and to the keys can facilitate or hinder the process of arriving at the pair of keys. For example, the extent to which distractor–key pairs occur in text in similar contexts without being synonyms in the context of the stem could attract the attention of test takers who have a less well-developed or deeper vocabulary and make the item harder, whereas to the extent that they do not share those contexts, they could be more easily seen as distractors and make the item easier.

Reliance on word familiarity is consistent with Lohman's (2000) explanation for the correlation between vocabulary and reasoning items. There is reason to believe that an individual's mental lexicon is the result of a process not unlike the vocabulary that results from the processing of corpora. That is, each person's lexicon can be seen as the result of processing a personalized corpus defined by the language the person has been exposed to, although the process is moderated by age, education, and multilingualism (Keuleers, Stevens, Mandera, & Brysbaert, 2015). Specifically, the psychological, as opposed to statistical, reality of word frequency is documented by evidence relating word frequency to lexical decision tasks (Keuleers et al., 2015), for example.

In short, although word familiarity is a reasonable component of sentence equivalence item difficulty, if it were the only component of difficulty, the contribution of the sentence equivalence item type to the GRE Verbal score could be questioned because it would not fulfill the design intentions of requiring a more complex response that mimics the reading and vocabulary acquisition process. To investigate whether this deeper process can be assumed to underlie the response process to sentence equivalence items, we evaluated a difficulty model containing indicators of the three subprocesses: context, word familiarity, and depth of word familiarity. That is, we first regressed difficulty on context features, followed by context and familiarity features, and finally adding depth of familiarity features. To the extent that these features reflect the item solutions process, we expect that incorporating them in the regression will result in a statistically significant increase in prediction of difficulty.

Method

Datasets

The 800 items analyzed in Studies 1 and 2 were split randomly into a 300-item development set and a 500-item test set.

Item Difficulty

We had available two measures of difficulty that are operationally maintained by the GRE program, the estimated IRT difficulty parameter, *b*, and an estimate of the proportion correct for a population. (Both of these are "equated" or put on a common metric.) IRT difficulty, *b*, was used in the analysis described subsequently to identify features for modeling difficulty. For purposes of the main modeling of difficulty by testing components of difficulty, we chose the proportion correct because it provides a more complete characterization of difficulty, namely, the probability that a randomly chosen test taker from the population will answer the item correctly, and was chosen as the dependent variable in the analyses. By contract, *b* does not have that population interpretation. Of course, the two measures of difficulty are strongly related, although not linearly. (The Pearson correlation on the 800 items is −.96.)

Feature Universe

There are a multitude of ways to characterize words and sentences. We took advantage of features that are under development at ETS to support a variety of assessments, especially the characterization of vocabulary in the *e-rater®* automated scoring engine. The features were modified for this project to characterize sentence equivalence items. A subset of features was chosen from exploratory analyses, including a principal components analysis of all the features to understand the latent structure among them. The original feature universe was generated by considering a range of base features, defined at the level of individual words or word pairs, item structure (is the word part of the stem, a key, or a distractor? Do word pairs relate stem to key, stem to distractor, key to key, or key to distractor?), and an aggregation rule (average, minimum, maximum). The resulting features thus had an implied structure. Some of the underlying linguistic measures (such as frequency, AofA, and word length) are expected to reflect a common underlying familiarity factor, whereas others (such as pointwise mutual information for *n*-grams and the mutual rank ratio statistic for *n*-grams) are expected to reflect a different underlying factor (formulaicity of language). We expected—and, in fact, reproduced—this implied structure in the exploratory principal components analysis. That is, features that addressed the same structural element (e.g., the key) with features expected to be related (such as word frequency and word length) loaded on the same components in the principal components analysis. This allowed us to consider aligned features as alternative measures of an item attribute (such as familiarity of the key, or the formulaicity of the relation between the key and its local context in the stem, or the strength of association between the key and the distractors).

Stepwise regression analysis with difficulty as a dependent variable was used as the means to identify a set of candidate features. For these analyses, *b* was used as a dependent variable in the stepwise regression. The development dataset was used for that purpose. Exploratory regressions were run separately for the components identified in the principal components analysis to identify features from each component that appeared to have the most significant associations with difficulty. This allowed us to identify a relatively small set of features for each potential construct attribute.

Based on this exploratory feature analysis, a set of eight features was chosen to characterize the language of the stem, or the context in which the vocabulary question is presented, the level of familiarity required by the keys, as well as the depth of familiarity required by the items, that is, the nature of the distractors with respect to the keys. Although stepwise regression is frowned upon because of its theoretical approach (Cohen, Cohen, West, & Aiken, 2003), in the present case, it was used for expediency, although guided by the nature of this investigation. Moreover, the model based on the development set was cross-validated on the development set of items that was not used in the identification of features, as described later.

Context Features

The following features were used to characterize the language of the stem and were identified empirically by means of stepwise regression, with one exception: the length or number of words in the stem.

The context features are as follows:

- *C1 (Stem Average of Average Trigram PMI)*. Characterizes the language in the stem or context. The computation of this feature is based on the point mutual information (PMI; Church & Hanks, 1989) statistic. Trigram PMI is computed as the ratio of the probability of the co-occurrence of each component word in a three-word sequence to the product of the probability of each word by itself:

$$\mathrm{PMI} = \log\left[\frac{P\left(x, y, z\right)}{P\left(x\right) P\left(y\right) P\left(z\right)}\right]$$

Intuitively, to the extent the words in a phrase are related, they would tend to occur together within the same phrase (Turney, 2001), more often than if the words are not related. In this study, we use trigram PMIs derived from a large corpus containing several combined corpora, including Sourcefinder (Passonneau, Hemat, Plante, & Sheehan, 2002) and Gigaword 2009 (Parker, Graff, Kong, Chen, & Maeda, 2009), a very large corpus (in excess of 2 billion words of newswire text), resulting in a very large aggregated corpus. As applied to sentence equivalence items, all the possible trigrams derived from the stem are extracted and the PMI for each trigram is calculated, after which the mean PMI across all trigrams in the stem is computed.

- *C2 (Key Average of Average Gigaword MRR)*. This feature is based on the mutual rank ratio (MRR; Deane, 2005), a statistic intended to characterize the tendency of a set of words that tend to appear together (i.e., as a “stock” phrase) in some corpus. The MRR statistic, like PMI, is a measure of how much more frequently a particular word sequence appears than would be expected by chance (for details of how it is calculated and an explanation of the statistical basis of the feature, see Deane, 2005). However, MRR values are defined in a fashion that yields values on the same scale for word sequences of multiple length (bigrams, trigrams) and so allow bigram and trigram information to be combined in a single summary measure. In this case, we use MRR values calculated from the Gigaword 2009 corpus. The MRR statistic is applied to keys as follows: First, we identify the two words that appear before the blank in the stem and the two words that appear after the blank then treat the key as having been inserted in the blank. That yields the sequence W1, W2, Key, W3, W4, which corresponds to two trigrams containing the key (W2 – Key, Key – W3) and three bigrams containing the key (W1 – W2 – Key, W2 – Key – W3, Key – W3 – W4). The average of these five MRR values yields the average MRR value for the key in that sentence context, which we then average across the pair of keys that appears in the item.
- *C3 (Distractor Average of Average Gigaword MRR)*. This feature is calculated exactly as described for C2, except that we use the four distractor words instead of the two keys to fill in the blank in the local sentence context in the stem.
- *C4*. Unlike the previous three features, which are computationally complex and were chosen based on the stepwise regression analysis, the fourth feature is length, or simply the number of words in the stem.

Familiarity Features

A range of different statistics can be used to characterize the familiarity of words, as discussed in Study 2. The variable that emerged from the exploratory analysis based on the development set of items was based on the AofA characterization of word familiarity. They are computed as follows:

- *F1 (Distractor Average of Average Age of Acquisition Mean)*. Retrieve the mean AofA for each word that appears as part of a distractor option. Then average this result across all four distractors.
- *F2 (Key Average of Min of Age of Acquisition Mean)*. Retrieve the mean AofA for each word that appears as part of a key. Then average this result across both keys.

Depth of Familiarity

Depth of familiarity refers to the relationship distractors and the keys as well as between keys. One feature, DF1, evaluates the similarity between the two keys based on latent semantic analysis (LSA; Landauer & Dumais, 1997). In this approach, words are represented as vectors in a multidimensional space. The cosine of the vectors is the estimate of the relationship between two words:

- *DF1 (Avg. Semantic Vector Cosine from KEY to KEY)*. Using a set of LSA vectors derived from the TASA corpus (Zeno et al., 1995), take all cosines between words in the first key to words in the second and average them.
- *DF2 (Min PMI from KEY to DISTRACTOR)*. This feature uses a different PMI calculation than described under heading C1. It is intended to capture the strength of association between the key and the distractor. The underlying statistic is calculated as

$$\mathrm{PMI} = \log\left[\frac{P\left(\text{key, distractor}\right)}{P\left(\text{key}\right)P\left(\text{distractor}\right)}\right]$$

Table 13 Prediction of Difficulty Based on Context Features (C) and Incremental Prediction due to Familiarity (F) and Depth of Familiarity (DF) for the Development Dataset

Model	R	R^2	Adj. R^2	SE	ΔR^2	ΔF	$df1$	$df2$	Δ sig. F
C	0.262	0.068	0.056	0.172	0.068	5.422	4	295	0.000
C + F	0.464	0.215	0.199	0.158	0.146	27.324	2	293	0.000
C + F + DF	0.511	0.261	0.241	0.154	0.047	9.181	2	291	0.000

Note. $n = 300$.

Table 14 Prediction of Difficulty Based on Context Features (C) and Incremental Prediction due to Familiarity (F) and Depth of Familiarity (DF) for the Test Dataset

Model	R	R^2	Adj. R^2	SE	ΔR^2	ΔF	$df1$	$df2$	Δ sig. F
C	0.166	0.028	0.020	0.179	0.028	3.523	4	495	0.008
C + F	0.339	0.115	0.104	0.171	0.087	24.292	2	493	0.000
C + F + DF	0.371	0.138	0.124	0.169	0.023	6.455	2	491	0.002

Note. $n = 500$.

where the underlying frequencies are *not* frequency of co-occurrence within an n-gram (as described in C1) but frequency of co-occurrence within the same paragraph window, using frequencies from a large corpus that includes the Sourcefinder and the Gigaword 2009 corpora. Once PMI is calculated for each distractor – key pair, the minimum of the eight possible PMIs is identified and returned as the value of this feature, which can be interpreted as providing a lower bound on the extent to which the key is related to (and hence potentially confusable with) the distractors.

Having defined blocks of variables, C1 – 4, F1 – 2, DF1 – 2, to characterize the context, familiarity, and depth of familiarity, respectively, we regressed difficulty incrementally by entering the blocks in that order and evaluating whether successive blocks incremented the prediction of difficulty variability significantly. (The specific order was chosen to correspond with the hypothesized order in which the test takers interacts with the item, as described earlier.) As it was the case with Study 2, there were missing data because the features could not be computed in every case. In such cases, we substituted the mean. This is far from ideal but is likely to be a conservative option. That is, substituting the mean is likely to attenuate the relationship among the predictors and possibly underestimate their predictive power, that is, by substituting the mean.[10]

Results

Table 13 shows the summary of the three corresponding models for the development dataset based on 300 items. Not surprisingly, because the features were identified in the development set, each of the three postulated components significantly increases prediction. That is, context, which was entered first, accounts significantly for difficulty variability on its own. The contribution of familiarity, partialing out context, is also significant. Finally, the contribution of depth of familiarity partialing out both context and familiarity is significant. The adjusted R^2 for the entire model with all three components is, approximately, 25% of the difficulty variance that is accounted.

Because the development set was used to identify features, the results need to be interpreted with a grain of salt, however, especially in light of the use of stepwise regression to identify features. To better evaluate the independent contribution of context, familiarity, and depth of familiarity, we validated the analysis of the development dataset consisting of 500 items that had not been used in the identification of features. The results are shown in Table 14. As can be seen, all three components are significant. The adjusted R^2 for the model with all components has shrunk to 12%, in contrast to the 25% accounted for by the model based on the development set.

Summary of Study 3

We evaluated a difficulty model for sentence equivalence items by postulating a response model that requires more than word familiarity, as may have been the case with earlier GRE item types. As discussed more fully later, the results are consistent with the notion that the response process required by the sentence equivalence item type requires deeper processing and therefore supports the presence of construct representation.

Conclusion and Limitations

The goals of the study were to examine the sentence equivalence item type introduced in the 2011 revision of the GRE. To that effect, we sought support for the generalization inference by describing the universe of items and the nature of the vocabulary tested. In addition, we sought evidence of construct representation or whether performance on the items requires more than word familiarity, as might have been the case with the earlier item types, and the appropriateness of the vocabulary tested given the purpose of the GRE.

In Study 1, we provided information about the structure of the items that led us to the conclusion that the templates from which items are produced are well defined and make it possible, in principle, to describe the universe of items. Access to such a description supports the generalization inference because we can characterize samples from the universe of items in terms of the POS of the blank and the possible semantic templates of the stem, as well as stem length, thereby partially defining the rules of inclusion in the universe.

Study 2 completed the description of the item universe by analyzing the vocabulary of the stem, the keys, and the distractors. The vocabulary for a graduate-level examination reasonably includes less frequent words. As a result, there were missing data, because the necessary information for those words was missing in some cases. That is, this is a case of nonignorable missing data because it is the less frequent words and *n*-grams that are affected. A comprehensive solution to this case of missing data is a major undertaking because there are no off-the-shelf solutions.[11] Despite the complexities introduced by missing data, we nevertheless conclude that the level word familiarity required by GRE appropriately spans a wide range of word frequencies, including some words that are rather infrequent. Perhaps the most compelling argument to support that conclusion is that the vocabulary of GRE sentence equivalence items was on par with the vocabulary GRE test takers deploy on their own when writing GRE essays.

Study 3 was primarily concerned with construct representation. The study sought to provide statistical evidence that the new vocabulary item type elicits vocabulary knowledge taking advantage of the context in which a word appears, in contrast to the earlier decontextualized vocabulary items, thereby being more in line with the design objective of eliciting vocabulary knowledge in a format that is closer to the vocabulary acquisition process and not subject to memorization. We assumed a response process that involves the context as well as word familiarity and depth of word familiarity and tested their independent contributions to the prediction of difficulty. All three components contributed to prediction in the test dataset, although the proportion of variance accounted for by the model was small. However, the goal was not to develop a comprehensive difficulty model at this point.

It is instructive to compare our results to difficulty modeling of GRE (Gorin & Embretson, 2006), the *TOEFL*® test (Sheehan, Ginther, & Schedl, 1999), and Armed Services Vocational and Aptitude Battery (ASVAB; Embretson & Wetzel, 1987) paragraph comprehension items, because in those cases, difficulty is modeled as a function of passage attributes and question attributes. That is, the passage corresponds to the "stem." Gorin and Embretson (2006) reported that the adjusted R^2 in each of those cases was .34 (see their Table 4) for GRE items, .25 for TOEFL items, and .28 for ASVAB items. The corresponding R^2 for sentence equivalence items, based on the development sample, was .25, which is of the same order of magnitude as those reported by Gorin and Embretson. However, when cross-validated on the test sample, the sentence equivalence adjusted R^2 dropped substantially, down to .12. We suspect that the lower estimates we obtained on the test sample are more realistic, in general. That is, there is potential for capitalizing on the idiosyncrasies of a sample of items when selecting features and when estimating the model, especially when there are relatively few items. In this investigation, we had access to a much larger item pool and could evaluate the model more realistically on a set of items that was not used in the development of the model.[12]

Apart from the total variance accounted for, the contribution of the different components weighs more heavily in a conclusion regarding construct representation. We found that the three postulated components of difficulty were significant, although familiarity and depth of familiarity contributed the bulk of the prediction of difficulty.

Granted that the goal was not to develop a model of difficulty that would achieve a high level of prediction, it is nevertheless informative to speculate what would be needed to do so and the implications of succeeding or not. A difficulty model can have a purely pragmatic goal of maximizing the prediction of difficulty to reduce or eliminate pretesting, for example. For pragmatic purposes, maximal prediction, however it is achieved, seems acceptable. For validity purposes, however, improving the level of prediction should be the result of construct-relevant features and a more detailed model of the item solution process. The model we proposed assumes that the sentence, the keys, and the distractors all have a role to play, but each could be elaborated along several dimensions. Specifically, the context features could include semantic attributes of the sentence, such as the taxonomy discussed in Study 1. By contrast, word familiarity and depth of familiarity, and the interrelationships among keys and distractors, could be sufficiently well described with existing features. Clearly the state of the art and resources for characterizing words are further ahead than the characterization of sentences. Further research on the contribution of context to the difficulty of sentence equivalence items would be useful toward a better understanding of the sentence equivalence item type.

Acknowledgments

We are grateful to Dan McCafree, who was the internal reviewer and offered many valuable suggestions on the statistical analysis. Two GRE TAC members, Jimmy de la Torre and Keena Arbuthnot, had many insightful methodological and substantive comments that have led to a significantly improved report. Finally, Adam Banta served as consultant on the sentence equivalence item type and patiently answered the many questions we posed to him.

Notes

[1] In this report, we do not present evidence concerning the extrapolation or decision inferences as they apply to an admissions test. The extrapolation inference in the case of an admissions test requires evidence that that score on the test is related to a relevant criterion, such as grades or predictive validity, and it is beyond the scope of this study. Similarly, the decision inference pertains to the full scores and how they are used in making admissions decisions, which is beyond the scope of the project. Nevertheless, because the evidence from multiple inferences is needed to formulate a validity argument for GRE scores, the results from this investigation contribute to that argument.

[2] Examples from Figures 1 and 2 can be found online at https://www.ets.org/gre/revised_general/prepare/verbal_reasoning/sentence_equivalence/sample_questions

[3] Examples of multiword expressions are “ferret out,” “stand up,” and so on. Multiword expressions present technical and conceptual challenges. The meaning of a multiword expression is not necessarily a function of the component words. Therefore they need to be treated as if the component words are words in their own right. However, in text, they occur as a sequence of words, which introduces significant processing complexities. For example, corpora would need to be specially prepared by tagging multiword expressions, which, in turn, requires the availability of an agreed-on list of multiword expressions. The problem is considered “vexing” by linguists and computational linguists. Multiword expressions are a current hot topic in linguistic and NLP research. On the positive side, there is reason to believe that, among native speakers, MWE are no different than regular words (Arnon & Snider, 2010). Conversely, MWE could be differentially difficult for nonnative speakers (Martinez & Murphy, 2011).

[4] The document’s title is “Development of a Revised Verbal Reasoning Measure for GRE: Summary of Design and Development.”

[5] The test developer who carried out this task reported that the task was not straightforward and that other test developers could easily classify the items differently.

[6] It should be noted that a strong association between difficulty and stem length could be problematic. For example, if length were strongly related to difficulty, it could trigger construct-irrelevant response strategies by test takers.

[7] The GRE is, of course, taken by students whose first language is not English. Score recipients, such as admissions offices, presumably take into account the language background of the applicant and make necessary adjustments in the interpretation of scores.

[8] SFI have the following interpretation: Lower numbers indicate lesser frequency. An SFI of 70 means it occurs once every 1,000 words of text; 60, every 10,000 words; 50, every 100,000 words; and 40, every 1 million words.

[9] Note, however, that we are referring to the pool as a whole. Because the lexical attributes we studied are not used in the assembly of forms, it cannot be assumed that every form has equivalent lexical characteristics.

[10] Features involving the Gigaword corpus, C2 and C3, had 12% missing data and were, by far, the worst case. For the remaining features, missing data ranged from 0% to 4%.

[11] The problem has a counterpart in health applications where patients leave for the very same health reason the study is concerned with (Schafer & Graham, 2002).

[12] Gorin and Embretson used IRT b as the dependent variable, whereas we used proportion correct. However, the fact that we used equated proportion correct as the dependent variable, rather than IRT b, is not the reason for the lower R^2 reported here: For the data used in this study, the R^2 with b as the dependent variable was the same.

References

Andrews, G., Birney, D., & Halford, G. S. (2006). Relational processing and working memory capacity in comprehension of relative clause sentences. *Memory & Cognition*, *34*, 1325–1340.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*, 67–82.

Bejar, I. I., Chaffin, R., & Embretson, S. E. (1991). *Cognitive and psychometric analysis of analogical problem solving*. New York, NY: Springer.

Bejar, I. I., Stabler, E. P., Jr., & Camp, R. (1987). *Syntactic complexity and psychometric difficulty: A preliminary investigation* (Research Report No. RR-87-25). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1987.tb00229.x.

Biemiller, A., & Boote, C. (2006). An effective method for building meaning vocabulary in primary grades. *Journal of Educational Psychology*, *98*(1), 44–62.

Braun, H., Kirsch, I., & Yamamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th-grade NAEP reading assessment. *Teachers College Record*, *113*, 2309–2344.

Breland, H. M. (1996). Word frequency and word difficulty: A comparison of counts in four corpora. *Psychological Science*, *7*(2), 96–99.

Breland, H. M., Jones, R. J., & Jenkins, L. (1994). *The College Board vocabulary study* (College Board Report No. 94-4; ETS Research Report No. RR-94-26). New York, NY: College Entrance Examination Board. http://dx.doi.org/10.1002/j.2333-8504.1994.tb01599.x.

Briel, J. B., & Michel, R. (2014). Revisiting the GRE general test. In C. Wendler & B. Bridgeman (Eds.), *The research foundation for the GRE® revised General Test: A compendium of studies*. Princeton, NJ: Educational testing Service. Retrieved from http://www.ets.org/s/research/pdf/gre_compendium.pdf

Burton, N. W., Welsh, C., Kostin, I., & VanEssen, T. (2009). *Toward a definition of verbal reasoning in higher education* (Research Report No. RR-09-33). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02190.x.

Carroll, J. B. (1980). Measurement of abilities constructs. *Construct validity in psychological measurement: Proceedings of a colloquium on theory and application in education and employment* (pp. 23–39). Princeton, NJ: Educational Testing Service.

Carroll, J. B., Davies, P., & Richman, B. (1971). *The American heritage word frequency book*. Boston, MA: Houghton Mifflin.

Carroll, J. B., & White, M. N. (1973). Word frequency and age of acquisition as determiners of picture-naming latency. *Quarterly Journal of Experimental Psychology*, *25*(1), 85–95.

Church, K. W., & Hanks, P. (1989). Word association norms, mutual information and lexicography. *Proceedings of the 27th Annual Conference of the Association of Computational Linguistics* (pp. 76–83). Strousdburg, PA: Association of Computational Linguistics.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Dale, E., & O'Rourke, J. (1976). *The Living Word Vocabulary, the words we know: A national vocabulary inventory*. Elgin, IL: Dome.

Deane, P. (2003). Co-occurrence and constructions. In L. Lagerwerf, W. Spooren, & L. Desgand (Eds.), *Determination of information and tenor in texts: Multidisciplinary approaches to discourse* (pp. 277–304). Amsterdam, Netherlands: Stichting Neerlandistiek.

Deane, P. (2005, June). *A nonparametric method for extraction of candidate phrasal terms.* Paper presented at the 43rd annual meeting of the Association for Computational Linguistics, Ann Arbor, MI. Retrieved from https://www.ets.org/Media/Research/pdf/erater_acl2005.pdf

Deane, P., Lawless, R., Li, C., Sabatini, J., Bejar, I. I., & O'Reilly, T. (2014). *Creating vocabulary item types that measure students' depth of semantic knowledge* (Research Report No. RR-14-02). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12001.

Deane, P., & Sheehan, K. M. (2003). *Automatic item generation via frame semantics: Natural language generation of math world problems*. Unpublished manuscript.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*, 179–197.

Embretson, S. E., & Gorin, J. S. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, *38*, 343–368.

Embretson, S. E., & Wetzel, D. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement*, *11*, 175–193.

Gorin, J. S., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement*, *30*, 394–411.

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*(1), 122–149.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Praeger.

Keuleers, E., Stevens, M., Mandera, P., & Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *Quarterly Journal of Experimental Psychology*, *68*, 1665–1692.

Kirkpatrick, J. J., & Cureton, E. E. (1949). Vocabulary item difficulty and word frequency. *Journal of Applied Psychology*, *33*, 347–351.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*, 978–990.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Lohman, D. F. (2000). Complex information processing and intelligence. In R. S. Sternberg (Ed.), *Handbook of intelligence* (pp. 285–340). Cambridge, England: Cambridge University Press.

Martinez, R. O. N., & Murphy, V. A. (2011). Effect of frequency and idiomaticity on second language reading comprehension. *TESOL Quarterly*, *45*, 267–290.

McKeown, M. G., & Curtis, M. E. (1987). *The nature of vocabulary acquisition*. Hillsdale, NJ: Lawrence Erlbaum.

Mislevy, R. J. (2011). *Evidence-centered design for simulation-based assessment* (CRESST Report No. 800). Retrieved from http://www.cse.ucla.edu/products/reports/R800.pdf

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *25*(4), 6–20.

Morrison, C. M., Ellis, A. W., & Quinlan, P. T. (1992). Age of acquisition, not word frequency, affects object naming, not object recognition. *Memory & Cognition*, *20*, 705–714.

Parker, R. C., Graff, D., Kong, J., Chen, K., & Maeda, K. (2009). Gigaword 9. In Linguistic Data Consortium, *English gigaword* (4th ed.). Philadelphia: University of Pennsylvania.

Passonneau, R., Hemat, L., Plante, J., & Sheehan, K. M. (2002). *Electronic sources as input to GRE reading comprehension item development: SourceFinder prototype evaluation* (Research Report No. RR-02-12). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2002.tb01879.x

Robin, F., Bejar, I. I., Liang, L., & Rijmen, F. (2016). *Dimensionality analyses of the GRE® revised General Test verbal and quantitative measures* (GRE Board Report No. 16-02). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12106.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177.

Shaoul, C., Harald Baayen, R., & Westbury, C. F. (2014). *N*-gram probability effects in a cloze task. *The Mental Lexicon*, *9*, 437–472.

Shaoul, C., Westbury, C. F., & Harald Baayen, R. (2013). The subjective frequency of word *n*-grams. *Psihologija*, *46*, 497–537.

Sheehan, K. M., Flor, M., & Napolitano, D. (2013, June). *A two-stage approach for generating unbiased estimates of text complexity*. Paper presented at the 2nd workshop of Natural Language Processing for Improving Textual Accessibility (NLP4ITA), Atlanta, GA.

Sheehan, K. M., Ginther, A., & Schedl, M. (1999). *Development of a proficiency scale for the TOEFL reading comprehension section*. Princeton, NJ: Educational Testing Service.

Sheehan, K. M., Kostin, I., & Futagi, Y. (2005). *A semi-automatic appraoch to assessing the verbal reasoning demands of GRE sentence completion items*. Unpublished manuscript.

Sheehan, K. M., & Mislevy, R. J. (2001). *An inquiry into the nature of the sentence-completion task: Implications for item generation* (Research Report No. RR-01-13). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2001.tb01855.x.

Turney, P. D. (2001). Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In L. Raedt & P. Flach (Eds.), *Lecture Notes in Computer Science: Vol. 2167 Machine Learning: ECML 2001* (pp. 491–502). Freiburg, Germany: Springer.

Wendler, C., & Bridgeman, B. (Eds.). (2014). *The research foundation for the GRE® revised General Test: A compendium of studies*. Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/s/research/pdf/gre_compendium.pdf

Zeno, S., Ivens, S. M., Koslin, S. H., & Zeno, B. L. (1995). *The educator's word frequency guide*. Brewster, NY: Touchstone Applied Science Associates.

Suggested citation:

Bejar, I. I., Deane, P. D., Flor, M., & Chen, J. (2016). *Evidence of the generalization and construct representation inferences for the GRE sentence equivalence item type* (GRE Board Report No. RR-17-02). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12134

Action Editor: Brent Bridgeman

Reviewers: This report was reviewed by the GRE Technical Advisory Committee and the Research Committee and Diversity, Equity and Inclusion Committee of the GRE Board.